A novel representation of protein sequences for prediction of subcellular location using support vector machines

SETSURO MATSUDA,[1] JEAN-PHILIPPE VERT,[2] HIROTO SAIGO,[1]
NOBUHISA UEDA,[1] HIROYUKI TOH,[3] AND TATSUYA AKUTSU[1]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0111, Japan
[2]Centre de Géostatistique, Ecole des Mines de Paris, 77300 Fontainebleau, France
[3]Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan

(RECEIVED May 20, 2005; FINAL REVISION August 22, 2005; ACCEPTED August 22, 2005)

Abstract

As the number of complete genomes rapidly increases, accurate methods to automatically predict the subcellular location of proteins are increasingly useful to help their functional annotation. In order to improve the predictive accuracy of the many prediction methods developed to date, a novel representation of protein sequences is proposed. This representation involves local compositions of amino acids and twin amino acids, and local frequencies of distance between successive (basic, hydrophobic, and other) amino acids. For calculating the local features, each sequence is split into three parts: N-terminal, middle, and C-terminal. The N-terminal part is further divided into four regions to consider ambiguity in the length and position of signal sequences. We tested this representation with support vector machines on two data sets extracted from the SWISS-PROT database. Through fivefold cross-validation tests, overall accuracies of more than 87% and 91% were obtained for eukaryotic and prokaryotic proteins, respectively. It is concluded that considering the respective features in the N-terminal, middle, and C-terminal parts is helpful to predict the subcellular location.

Keywords: subcellular location; signal sequence; amino acid composition; distance frequency; support vector machine; predictive accuracy

Predicting the subcellular location of proteins is important to infer their biological function. As the number of complete genomes rapidly increases, accurate methods that automatically predict the subcellular location become more necessary. In particular, in the case where no homologous protein is found in protein databases, such methods are important tools to help annotate the function of unknown proteins.

Reprint requests to: Setsuro Matsuda, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan; e-mail: smatsuda@kuicr.kyoto-u.ac.jp; fax: +81-774-38-3022.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051597405.

Many efforts have been made to develop prediction methods to date. PSORT (Nakai and Kanehisa 1992; Horton and Nakai 1997) is historically the first method for predicting subcellular locations. It uses various sequence-derived features such as the presence of sequence motifs and amino acid compositions. Most existing methods can be roughly classified into two groups according to their input data. One is the method based on the N-terminal sequence of a protein and the other on its amino acid composition. TargetP (Emanuelsson et al. 2000) requires the N-terminal sequence as an input into two layers of artificial neural networks (ANNs), and can also predict the peptidase-cleaved site of a protein. The first layer comprises the earlier binary predictors, SignalP (Nielsen et al. 1997) and ChloroP (Emanuelsson et al. 1999). Reczko and Hatzigeorgiou

(2004) used a bidirectional recurrent neural network with the first 90 residues in the N-terminal sequence. Yuan (1999) applied the Markov chain model to the prediction, but the entire sequence was used as the input data.

ProtLock (Cedano et al. 1997) requires the amino acid composition and is based on the least Mahalanobis distance algorithm. Chou and Elrod (1998, 1999) also used the amino acid composition but the covariant discriminant algorithm was employed in their method. NNPSL (Reinhardt and Hubbard 1998) is an ANN-based method using the amino acid composition. After the successful report in Reinhardt and Hubbard (1998), application of machine learning techniques became popular in this field. For SubLoc (Hua and Sun 2001), a support vector machine (SVM) was implemented instead of the ANN. It is expected that incorporating an amino acid order as well as the amino acid composition makes it possible to improve prediction performance. Chou (2001) proposed the pseudo–amino acid composition to take the effect of the amino acid order into account. Furthermore, Cai and Chou (2004) have recently developed an accurate method integrating the pseudo-amino acid composition, the functional domain composition (Chou and Cai 2002, 2004), and the information of gene ontology (Chou and Cai 2003). Park and Kanehisa (2003) developed an SVM-based method that incorporates compositions of dipeptides and gapped amino acid pairs in addition to the conventional amino acid composition. The concepts of the pseudo–amino acid and gapped amino acid pair compositions were merged in the residue-couple model proposed by Guo et al. (2005).

Incorporating the information of homology search can also improve the prediction performance (Bhasin and Raghava 2004; Kim et al. 2004; Bhasin et al. 2005). However, one should pay much attention to the sequence similarity between training and test data in evaluating prediction methods based on homology search. If a query sequence in the test data has a high similarity with a sequence in the training data, then its subcellular location can be easily predicted without using a complicated predictor. In other words, the data set used for training and testing must be sufficiently redundancy-reduced.

Although Reinhardt and Hubbard (1998) pointed out that prediction methods based on the amino acid composition are robust to the gene annotation error in the 5′-region, using the amino acid composition only leads to information loss of signal sequences. To overcome this problem, the concepts such as the pseudo–amino acid composition have been introduced. In this work, we propose a novel representation of protein sequences to further improve the accuracy of prediction methods. Our method, which employs the SVM with RBF kernel, is based on local compositions of amino acids and twin amino acids, and local frequencies of distance between successive amino acids. As benchmark data, we adopt the data sets provided by Reinhardt and Hubbard (1998) and Emanuelsson et al. (2000) because they have been widely used in earlier studies. For convenience, we call the former "NNPSL data sets" and the latter "TargetP data sets."

Each amino acid is represented by its one-letter code hereafter. In this work, basic amino acids encompass R, K, and H. Hydrophobic amino acids are I, V, L, F, M, A, G, W, and P. The remainder, D, N, E, Q, Y, S, T, and C are called "other amino acids."

Results

Tables 1 and 2 show the comparison of predictive accuracies with existing methods on the NNPSL data sets. Likewise, Tables 3 and 4 show that on the TargetP data sets. The overall accuracies of Chou and Cai (2003, 2004) are remarkably high for all data sets. The sensitivity, specificity, and MCC of their methods are not shown because these values are not given in their papers. As explained in the previous section, their methods require the information of gene ontology and functional domain retrieved from the InterPro database (Apweiler et al. 2001). On the contrary, all other methods use sequence information alone. Therefore, we cannot compare their methods with the other methods directly. It should be pointed out, however, that our protein representation may be incorporated into their methods.

In Table 1, our sensitivity for mitochondrial proteins is 0.13–0.25 higher than the other four methods. Since the average MCC of our method ($=0.82$) is the highest, it is clear that the performance of our method is well-balanced. In Table 2, the overall accuracy of our method is close to that of Guo et al. (2005). Although the performance of machine learning techniques such as ANN and SVM is affected by the number of training data, it seems that the discrimination between cytoplasmic and the remaining (Extra and Peri) proteins is relatively easy.

In Tables 3 and 4, our overall accuracies are the highest if we consider the jackknife accuracies of Chou and Cai (2004). According to Chou and Zhang (1995), the jackknife test is more rigorous and objective than cross-validation test, because the number of possible data divisions is too large to be handled in the latter test. However, we adopted the cross-validation test to save CPU time and compare our method with as many recent methods as possible. For plant proteins, our sensitivities for chloroplast, nuclear and cytosolic (other) proteins are lower than those of Emanuelsson et al. (2000), but the sensitivity for mitochondrial proteins was improved by 0.104. For non-plant proteins, our sensitivity for nuclear and cytosolic proteins is higher than any other

Table 1. *Comparison of predictive accuracies for eukaryotic proteins in the NNPSL data set*

Predictor	Location	Sensitivity	Specificity	MCC	Overall accuracy	Validation type
Our method	Cyto	0.825	0.812	0.75	0.871	Fivefold cross-validation
	Extra	0.892	0.942	0.90		
	Mito	0.819	0.858	0.81		
	Nuclear	0.909	0.893	0.82		
Guo et al. (2005)	Cyto	0.858		0.77	0.869	Fivefold cross-validation
	Extra	0.859		0.89		
	Mito	0.654		0.72		
	Nuclear	0.942		0.85		
Hua and Sun (2001)	Cyto	0.769		0.64	0.794	Jackknife
	Extra	0.800		0.78		
	Mito	0.567		0.58		
	Nuclear	0.874		0.75		
Yuan (1999)	Cyto	0.781		0.60	0.730	Jackknife
	Extra	0.622		0.63		
	Mito	0.692		0.53		
	Nuclear	0.741		0.68		
Reinhardt and Hubbard (1998)	Cyto	0.55			0.66	Independent data test
	Extra	0.75				
	Mito	0.61				
	Nuclear	0.72				
Chou and Cai (2003)					0.929	Jackknife

Cyto, Extra, Mito, and Nuclear indicate proteins destined for cytoplasm, extracell, mitochondria, and nucleus, respectively.

methods. It is noteworthy that the MCCs of our method are over 0.82 for all locations.

To compare the predictive accuracies in the same conditions, we implemented the method proposed by Kim et al. (2004). Therefore, the values of sensitivity, specificity, MCC, and overall accuracy are different from those in Kim et al. (2004). They also employed the SVM with RBF kernel and characterized protein sequences by the Needleman-Wunsch scores (Needleman and Wunsch 1970) against all the sequences in training data. ALIGN0 (Myers and Miller 1988) in the FASTA 2.0 package (Pearson and Lipman 1988; Pearson 1990) was used for calculating the scores. The gap penalty is −3 and the scoring matrix is BLOSUM50. Each sequence was truncated after the N-terminal 90 residues for the calculation. The values of

Table 2. *Comparison of predictive accuracies for prokaryotic proteins in the NNPSL data set*

Predictor	Location	Sensitivity	Specificity	MCC	Overall accuracy	Validation type
Our method	Cyto	0.985	0.938	0.87	0.917	Fivefold cross-validation
	Extra	0.737	0.873	0.78		
	Peri	0.777	0.863	0.78		
Guo et al. (2005)	Cyto	0.990		0.89	0.920	Fivefold cross-validation
	Extra	0.757		0.79		
	Peri	0.776		0.78		
Hua and Sun (2001)	Cyto	0.975		0.86	0.914	Jackknife
	Extra	0.766		0.77		
	Peri	0.782		0.78		
Yuan (1999)	Cyto	0.936		0.83	0.891	Jackknife
	Extra	0.776		0.77		
	Peri	0.797		0.69		
Chou and Elrod (1998)	Cyto	0.916			0.865	Jackknife
	Extra	0.804				
	Peri	0.723				
Reinhardt and Hubbard (1998)	Cyto	0.80			0.81	Independent data test
	Extra	0.77				
	Peri	0.85				
Chou and Cai (2003)					0.947	Jackknife

Cyto, Extra, and Peri indicate proteins destined for cytoplasm, extracell, and periplasm, respectively.

Table 3. *Comparison of predictive accuracies for plant proteins in the TargetP data set*

Predictor	Location	Sensitivity	Specificity	MCC	Overall accuracy	Validation type
Our method	cTP	0.7591	0.8474	0.7694	0.8809	Fivefold cross-validation
	mTP	0.9240	0.8652	0.8227		
	SP	0.9219	0.9326	0.8983		
	other	0.8210	0.8586	0.8070		
Kim et al. (2004)	cTP	0.6874	0.8435	0.7222	0.8479	Fivefold cross-validation
	mTP	0.8970	0.8392	0.7773		
	SP	0.8952	0.9428	0.8872		
	other	0.8027	0.7549	0.7296		
Emanuelsson et al. (2000)	cTP	0.85	0.69	0.72	0.853	Fivefold cross-validation
	mTP	0.82	0.90	0.77		
	SP	0.91	0.95	0.90		
	other	0.85	0.78	0.77		
Chou and Cai (2004)					0.923	Fivefold cross-validation
					0.854	Jackknife

cTP, mTP, SP, and "other" indicate proteins destined for chloroplast, mitochondria, secretory pathway, and other locations (nucleus and cytosol), respectively.

regularization parameter C and parameter γ of RBF kernel are the same as those in Kim et al. (2004; see Table 8, below).

Discussion

We proposed a new representation for protein sequences using distance frequencies of basic, hydrophobic, and other amino acids and separated a protein sequence into several regions. In this section, we discuss how the distance frequency is useful and whether the separation of a sequence is meaningful. Furthermore, we estimate and analyze the weights of features used in the representation.

Usefulness of the distance frequency

The distance frequency was developed in consideration of nuclear export signal (NES) and chloroplast transit peptide. In Figure 1, we visualized distance frequencies of three hydrophobic amino acids (L, I, and V) for 75 protein sequences containing the NES (called "with NES"). These sequences were downloaded from NES-base 1.0 (la Cour et al. 2003) and their NESs were experimentally verified. We also depicted the distance frequencies for the 75 sequences with their NES removed (called "without NES"). These frequencies are slightly smaller than those of "with NES" at $H = 2, 3, 4$, where H represents the distance between successive amino

Table 4. *Comparison of predictive accuracies for non-plant proteins in the TargetP data set*

Predictor	Location	Sensitivity	Specificity	MCC	Overall accuracy	Validation type
Our method	mTP	0.8303	0.8635	0.8228	0.9229	Fivefold cross-validation
	SP	0.9091	0.9118	0.8788		
	other	0.9498	0.9409	0.8609		
Kim et al. (2004)	mTP	0.6483	0.8569	0.7121	0.8762	Fivefold cross-validation
	SP	0.8530	0.8736	0.8158		
	other	0.9389	0.8819	0.7626		
Emanuelsson et al. (2000)	mTP	0.89	0.67	0.73	0.900	Fivefold cross-validation
	SP	0.96	0.92	0.92		
	other	0.88	0.97	0.82		
Reczko and Hatzigeorgiou (2004)	mTP	0.78	0.82	0.77	0.913	Fivefold cross-validation
	SP	0.93	0.91	0.89		
	other	0.93	0.94	0.84		
Chou and Cai (2004)					0.983	Fivefold cross-validation
					0.919	Jackknife

mTP, SP, and "other" indicate proteins destined for mitochondria, secretory pathway, and other locations (nucleus and cytosol), respectively.

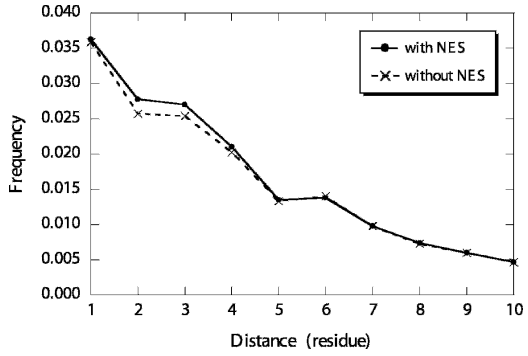

Figure 1. Distance frequencies of three hydrophobic amino acids (L, I, and V) for 75 protein sequences containing the NES (with NES). The dotted line shows the distance frequencies for the 75 sequences with their NES removed (without NES). Each value of the frequency was divided by sequence length and averaged over the 75 sequences.

acids. This decline implies that the distance frequency modestly reflects the existence of NES.

Distance frequencies for plant proteins in the TargetP data set are shown in Figure 2. Figure 2, A and B, represents distance frequencies of basic amino acids in the N-terminal and middle parts, respectively. Figure 2, C and D, represents those of hydrophobic and other amino acids in the middle part. In Figure 2A, the distance frequency for mTP is the largest and that for SP is the smallest when $1 < H \leq 6$. The difference of distance frequencies related to the two locations is significantly large compared with Figure 2B. This indicates that the distance frequency is useful for discrimination of mitochondrial and secretory proteins. In Figure 2, C and D, the difference of the frequencies between four locations seems to be small. However, the performance of our method was improved by incorporating the frequencies of hydrophobic and other amino acids.

Implication of separating a protein sequence

Each sequence was separated into three parts: N-terminal, middle, and C-terminal. The N-terminal part was further divided into four regions in calculating local amino acid compositions. Each region has 20 residues, except for prokaryotic proteins (24 residues). We tested different region lengths in the range 19–25. Interestingly, the highest overall accuracy was always obtained by using

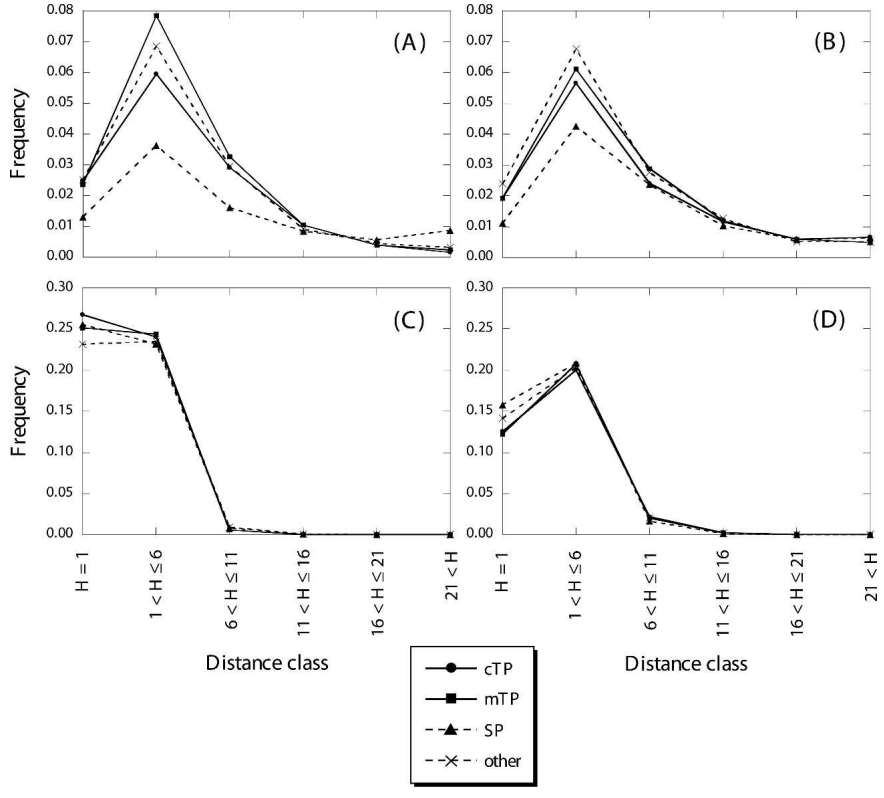

Figure 2. Distance frequencies of basic amino acids in the N-terminal part (*A*), basic amino acids in the middle part (*B*), hydrophobic amino acids in the middle part (*C*), and other amino acids in the middle part (*D*), for the TargetP plant proteins. Each value of the frequency was divided by sequence length and averaged over all sequences belonging to each subcellular location. The *X*-axis is common to the four panels. cTP, mTP, SP, and "other" indicate proteins destined for chloroplast, mitochondria, secretory pathway, and other locations (nucleus and cytosol), respectively.

20 residues. As a transmembrane domain consists of ~20 hydrophobic residues, the 20-residue length may have a biological meaning. For the C-terminal part, we changed the number of residues from 6 to 10. We observed that nine and eight residues were suitable on the NNPSL and TargetP data sets, respectively. The overall accuracy drastically varied depending on the number of residues in the C-terminal part. This indicates that considering the amino acid composition in the C terminus is important to predict the subcellular location. Although peroxisomal proteins, which can have the SKL motif in their C terminus, are not handled in this work, our method would be able to capture the peroxisomal targeting signal. We also examined two cases where the N-terminal part is divided into three regions with 30 residues and not divided at all. However, the overall accuracies were lower than the case mentioned above (four regions with 20 residues). As for the SVM parameters, we found that the effect of tuning them is not so critical compared with the adjustment of sequence lengths of the N-terminal and C-terminal parts.

Internal signal sequences, which are positioned in the middle part, are unclear compared with ones in the N-terminal and C-terminal parts. But some biological experiments indicate the importance of signal sequences in the middle part. Miyakawa and Imamura (2003) found out that two fibroblast growth factors FGF-9 and FGF-16 require both the N-terminal region and central hydrophobic region as a secretory signal. This hydrophobic region belongs to the middle part here. Furthermore, this bipartite signal sequence is not cleaved off by proteases during the transport process. We collected three sequences: human FGF-9, human FGF-16, and rat FGF-16 from databases available on the Internet and then predicted their subcellular locations by SignalP 3.0 (Bendtsen et al. 2004). This is the latest version of SignalP and employs both the ANN and hidden Markov model. As a result, these sequences were predicted as nonsecretory proteins. In contrast our method, which can consider the features in the middle part, correctly predicted all the sequences as secretory proteins.

From the aforementioned fact, it is concluded that separating a protein sequence into the N-terminal, middle, and C-terminal parts is helpful to capture signal sequences. In addition, our method has an advantage of small CPU time requirement to construct the feature vector compared with the method proposed by Kim et al. (2004).

Feature weights

Here we describe how to estimate the importance of each feature and discuss the relation between these features and subcellular locations. As opposed to linear SVM, the RBF SVM does not assign a weight to each feature. In order to estimate the importance of each feature, we followed the following procedures: (1) Prepare a feature vector whose components are all 0, (2) assign 1 to a feature whose importance is to be estimated, (3) feed this vector into the trained SVMs and obtain their outputs, and (4) repeat the procedures 1–3 for all features. The outputs are regarded as the weights of the RBF SVM, quantifying the contributions of the features.

Since our prediction method adopted the one-versus-rest method, we have one specific SVM for each subcellular location. Figure 3 shows the feature weights of the SVMs specifically for (A) SP and (B) "other" on the TargetP plant data set. Feature number j of the X-axis corresponds to the j-th component of a feature vector (see Equation 1). For easy understanding, we discuss the possible meaning of the features with the most positive weights. In Figure 3A, we can see that the weights of hydrophobic amino acids in the N-terminal 20 residues are large. Interestingly, the weights of cysteine in the N-terminal part are relatively large. It is noteworthy that the distance frequency of other amino acids in the middle part ($h_1^{(M)}$) has a large weight. In Figure 3B, it is clarified that aspartic and glutamic acids in the N-terminal 40 residues are important. We can also see that the weights of lysine in the N-terminal 20 residues and the middle part are large.

With respect to the SVM for cTP, it was confirmed that serine and threonine in the N-terminal 20 residues are strongly weighed. As to that for mTP, the weight of arginine in the N-terminal 20 residues was solely positive.

The above results indicate that the SVMs in our method were successfully trained, because their feature weights are consistent with features of signal sequences described later. Moreover, it is concluded that the first 20 residues in the N terminus are particularly important to predict the subcellular location.

Materials and methods

Data sets

In this work, the data sets provided by Reinhardt and Hubbard (1998) and Emanuelsson et al. (2000) were used. Both data sets (the NNPSL and TargetP data sets) were collected from the SWISS-PROT database and any sequences containing ambiguous residues such as X and B were excluded out of them. The NNPSL data sets (Table 5) consist of eukaryotic and prokaryotic proteins but do not include plant proteins. Within each subcellular location, none of the sequences has more than 90% identity to any other sequences. This criterion to reduce the redundancy is not strict, because we found that the subcellular locations except for mitochondria and periplasm can be predicted with a high accuracy (>82%) by simple homology search using the Smith-Waterman algorithm (Smith and Waterman 1981).

The TargetP data sets are comprised of two sets: plant and non-plant proteins (Table 6). However, the mitochondrial proteins contain sequences from both plant and non-plant

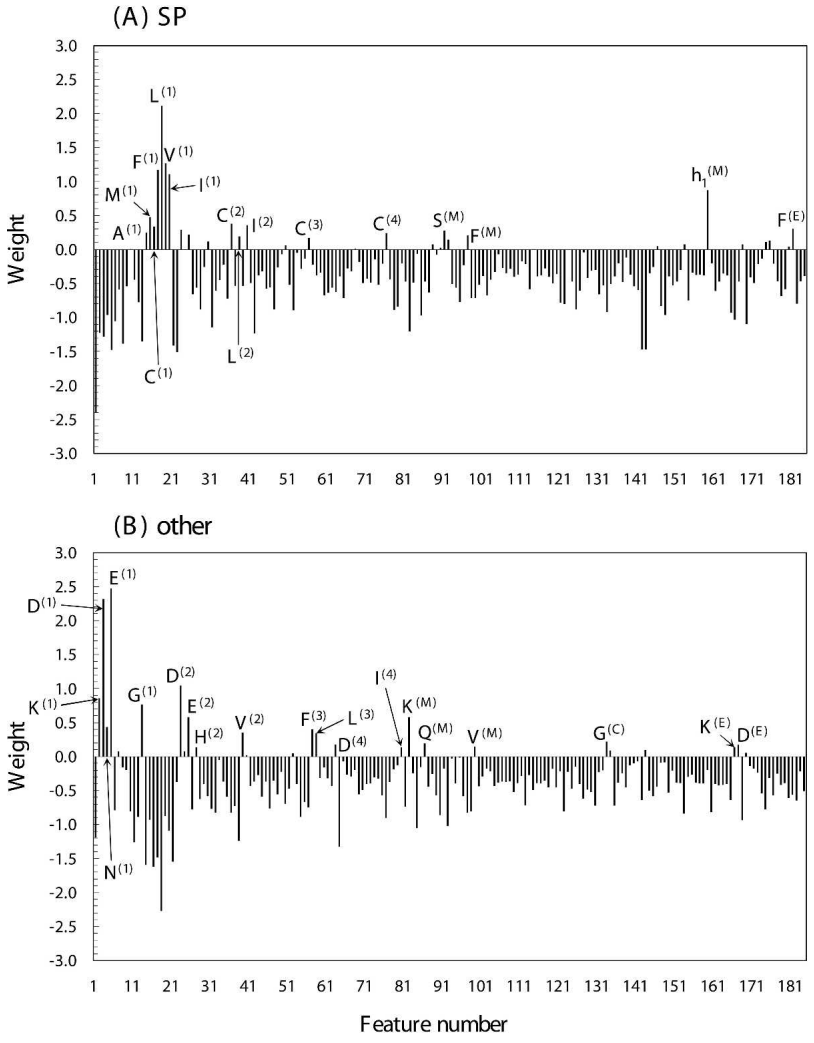

Figure 3. Feature weights of the SVMs specifically for SP (*A*) and "other" (*B*) on the TargetP plant data set. Feature number *j* of the *X*-axis corresponds to the *j*-th component of a feature vector. The capital letters represent amino acids and the superscripts indicate a region in a protein sequence. Refer to the definitions of the regions in Figure 4. $h_1^{(M)}$ represents the distance frequency of other amino acids in the middle part.

proteins, because the number of mitochondrial proteins extracted from SWISS-PROT was too small to be used. The redundancy reduction for plant proteins in cTP, mTP, and SP and for non-plant proteins in SP was done on their presequence plus the first residue of mature protein. Plant and non-plant proteins in "other" were redundancy-reduced on their N-terminal 68 residues. The remainder, non-plant proteins in mTP were redundancy-reduced on the mitochondrial targeting peptide plus three residues. To check the effect of the redundancy reduction, we predicted the subcellular locations based on the Smith-Waterman score. That is, the location of a sequence in the training data with a highest score is assigned to the corresponding query sequence in the test data. As a result, we obtained the overall accuracies of 75.7% and 84.0% for plant and non-plant proteins, respectively. This indicates that the redundancy on the TargetP data sets is relatively small.

In order to perform a fivefold cross-validation test, each data set was partitioned into five subsets that have approximately equal sizes. Before partitioning, we shuffled the sequences within each set by using at least 1000 random numbers. One subset is regarded as test data and the remaining four subsets as training data. This procedure is repeated five times so that each subset is used as test data once.

Important features of signal sequences

In general, proteins destined for chloroplast, mitochondria, and secretory pathway have signal sequences in their N termini. On the other hand, proteins destined for nucleus and cytosol have one or more signal sequences in the middle part of their sequence. Furthermore, chloroplast proteins transported into thylakoid have an internal signal sequence after the chloroplast transit peptide (cTP) (Keegstra and Cline 1999; Robinson et al. 2001).

The length of cTP is believed to be at most 100 residues. That of mitochondrial targeting peptide (mTP) ranges from 10 to 80

Table 5. *Number of sequences in each subcellular location on the NNPSL data sets*

Location	Eukaryotic (no. of sequences)	Location	Prokaryotic (no. of sequences)
Cytoplasmic (Cyto)	684	Cytoplasmic (Cyto)	688
Extracellular (Extra)	325	Extracellular (Extra)	107
Mitochondrial (Mito)	321	Periplasmic (Peri)	202
Nuclear	1097		—
Total	2427	Total	997

residues (Neupert 1997; Omura 1998). cTPs are rich in hydroxylated amino acids (S and T) and have basic amino acids with several residue gaps intervening (Bruce 2000). mTPs especially for mitochondrial matrix and intermembrane space can form amphipathic α-helix with basic amino acids (Omura 1998). Signal peptides (SPs) for secretion are abundant in hydrophobic amino acids (von Heijne 1990). Secretory proteins that have the KDEL or KKXX motif in their C terminus return from Golgi apparatus to endoplasmic reticulum (Cosson and Letourneur 1997).

The nuclear localization signal (NLS) and nuclear export signal (NES) are rich in basic and hydrophobic amino acids (particularly L, I, and V), respectively. The basic amino acids in NLS can form one or more clusters. NESs have the hydrophobic amino acids with approximately constant gaps between each hydrophobic amino acid. Some examples of signal sequences are summarized in Table 7.

Although sequence motifs such as the above were clarified by biological experiments, consensus sequences as localization signals are still obscure. This indicates that prediction of the subcellular location should not depend much on motif finding. As stated in the introduction of this paper, prediction methods based on the amino acid composition only take into account the whole length of a sequence and the methods based on the N-terminal sequence ignore the existence of signal sequences in the middle and C-terminal parts. Therefore, it would be effective that the three parts: N-terminal, middle, and C-terminal are separately treated to characterize protein sequences.

Feature vector

First of all, we defined the N-terminal, middle, and C-terminal parts depending on sequence length L. Most of the sequences used here conform to the definition in Figure 4A. The N-terminal part is further divided into four regions with length d_N. Because we assumed that proteins are directed by the approximate amount of specific amino acids to make the signal sequence flexible and the cluster of such amino acids can be distributed in various regions even in the N terminus. d_N is set to 20 and 24 for eukaryotic and prokaryotic proteins, respectively. It was also assumed that the middle part has at least 20 residues equal to the number of distinct amino acids. The length of the C-terminal part d_C is set to nine and eight on the NNPSL and TargetP data sets, respectively.

For short sequences, we prepared two more definitions. If L is $> 4d_N + d_C$ and $< 4d_N + 20 + d_C$, the middle part is regarded as 20 residues from the start of the C-terminal part toward the N-terminal part (Fig. 4B). In the case that L is $\leq 4d_N + d_C$, we assumed that the lengths of the N-terminal and middle parts are the same. That is, these lengths are defined by $(L - d_C)/2$ and the N-terminal part is not divided at all (Fig. 4C). Actually, the sequences that satisfy $L < 4d_N + 20 + d_C$ are only 3.7%–6.3% of the data sets.

The feature vector to represent protein sequence i is expressed as follows:

$$\begin{aligned}\boldsymbol{v}_i = (&x_1^{(1)}, ..., x_{20}^{(1)}, x_1^{(2)}, ..., x_{20}^{(2)}, x_1^{(3)}, ..., x_{20}^{(3)}, x_1^{(4)}, ..., x_{20}^{(4)},\\ &x_1^{(\mathrm{M})}, ..., x_{20}^{(\mathrm{M})}, y_1^{(\mathrm{M})}, ..., y_{20}^{(\mathrm{M})}, x_1^{(\mathrm{C})}, ..., x_{20}^{(\mathrm{C})}, f_1^{(\mathrm{N})}, ..., f_6^{(\mathrm{N})},\\ &f_1^{(\mathrm{M})}, ..., f_6^{(\mathrm{M})}, g_1^{(\mathrm{M})}, ..., g_6^{(\mathrm{M})}, h_1^{(\mathrm{M})}, ..., h_6^{(\mathrm{M})}, x_1^{(\mathrm{E})}, ..., x_{20}^{(\mathrm{E})})^{\mathrm{T}},\end{aligned} \quad (1)$$

where the capital letters, N, M, C, and E indicate the N-terminal, middle, C-terminal, and entire parts. The entire part means the whole length of a sequence. The numerals in the parentheses (1–4) correspond to the regions in the N-terminal part in Figure 4, A and B. $x_1^{(p)}, ..., x_{20}^{(p)}$ indicate the composition of 20 amino acids in part p (p = 1, 2, 3, 4, M, C, E). $y_1^{(\mathrm{M})}, ..., y_{20}^{(\mathrm{M})}$ are the composition of 20 twin amino acids (e.g., RR, KK) in the middle part. In the case that a sequence is too short to be divided on its N-terminal part (see Fig. 4C), the amino acid composition of the whole N-terminal residues is equally assigned to the four regions, i.e., $x_j^{(1)} = x_j^{(2)} = x_j^{(3)} = x_j^{(4)}$ (j = 1, ..., 20).

$f_1^{(q)}, ..., f_6^{(q)}$ represent the distance frequencies of basic amino acids in part q (q = N, M). To calculate distance frequencies, we defined six distance classes (H = 1, $1 < H \leq 6$, $6 < H \leq 11$, $11 < H \leq 16$, $16 < H \leq 21$, $H > 21$). Similarly, $g_1^{(\mathrm{M})}, ..., g_6^{(\mathrm{M})}$ are the distance frequencies of hydrophobic amino acids and $h_1^{(\mathrm{M})}, ..., h_6^{(\mathrm{M})}$ are those of other amino acids in the middle part. Altogether, this feature vector has 184 components and each component is normalized between 0 and 1 by its possible maximum.

Distance frequency

In this work, we introduced a new feature, called "distance frequency" to encode a protein sequence. This is the frequency of the distance between two successive amino acids. For example, consider the following protein sequence:

AAKAARARAAKAKAAHA,

where underlined letters denote basic amino acids. The distances between successive basic amino acids, H_b, take the values 3, 2, 3, 2, and 3 starting from the left. Note that H_b is calculated in a left-to-right fashion. As a result, the distance frequencies for H_b = 2 and H_b = 3 are 2 and 3, respectively.

Table 6. *Number of sequences in each subcellular location on the TargetP data sets*

	No. of sequences	
Subcellular location	Plant	Non-plant
Chloroplast (cTP)	141	—
Mitochondrial (mTP)	368	371
Secretory (SP)	269	715
Nuclear + cytosolic (other)	162	1652
Total	940	2738

Table 7. *Signal sequences and their target locations*

Function of signal sequence	Example of signal sequence
Import into chloroplast	NH_2 -MVAMAMASLQSSMSSLSLSSNSFLGQPLSPITLSPFLQG-
	NH_2 -MAMAMRSTFAARVGAPAVRGARPASRMSCMA-
Import into mitochondria	NH_2 -MLSLRQSIRFFKPATRTLCSSRYLL-
Import into ER	NH_2 -MMSFVSLLLVGILFWATEAEQLTKCEVFQ-
Return to ER	-KDEL-COOH (KDEL motif)
	-KKXX-COOH (dilysine motif)
Import into nucleus	-PKKKRKV- (single type)
	-RQARRNRRRWE- (arginine-rich type)
	-KRPAAIKKAGQAKKKK- (bipartite type)
	-NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY-
Export from nucleus	-LALKLAGLDL- (leucine-rich type)

NH_2 and COOH indicate the N terminus and C terminus of a protein. Important amino acids for the function of the signal sequence are underlined. ER is the abbreviation of endoplasmic reticulum and X represents an arbitrary amino acid.

SVM training

In order to implement SVM, we used the free software, SVM^{light} developed by Joachims (1999). As the kernel, the radial basis function (RBF) was selected because this function outperformed linear and polynomial kernels in terms of overall predictive accuracy (data not shown). The RBF kernel is defined by the following equation:

$$K(\boldsymbol{v}_i, \boldsymbol{v}_j) = \exp\left(-\gamma \left\| \boldsymbol{v}_i - \boldsymbol{v}_j \right\|^2\right), \quad (2)$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are feature vectors representing protein sequences. The parameter γ in Equation 2 and regularization parameter C are adjusted in training to produce reliable performance. As γ becomes smaller, the decision boundary for discriminating positive and negative examples becomes smoother. C controls the trade-off between training error and margin. We determined the two parameters as shown in Table 8 by trial and error. Other options for SVM^{light} are set to their default.

For multiclass classification, the one-versus-rest method (Schölkopf and Smola 2002; Nguyen and Rajapakse 2003) was adopted. That is, the l-th SVM is trained on sequences belonging to the l-th location with the positive label "+1" and on sequences belonging to the remaining locations with the negative label "−1." We also tested the one-versus-one method, but the overall accuracy was lower than the one-versus-rest method (data not shown).

Measures for evaluation of the prediction performance

To evaluate the prediction performance of our method, sensitivity, specificity, Matthews' (1975) correlation coefficient (MCC) for each subcellular location, and overall accuracy were calculated. The definitions of these measures are as follows:

$$Sensitivity(l) = \frac{tp(l)}{tp(l) + fn(l)}, \quad (3)$$

$$Specificity(l) = \frac{tp(l)}{tp(l) + fp(l)}, \quad (4)$$

$$MCC(l) = \frac{tp(l) \times tn(l) - fp(l) \times fn(l)}{\sqrt{(tp(l) + fn(l))(tp(l) + fp(l))(tn(l) + fp(l))(tn(l) + fn(l))}}, \quad (5)$$

$$Overall\ accuracy = \frac{1}{n} \sum_{l=1}^{k} tp(l), \quad (6)$$

where n is the total number of protein sequences and k is the number of subcellular locations. $tp(l)$ is the number of correctly predicted sequences belonging to location l (true positive). $tn(l)$ is the number of correctly predicted sequences that do not belong

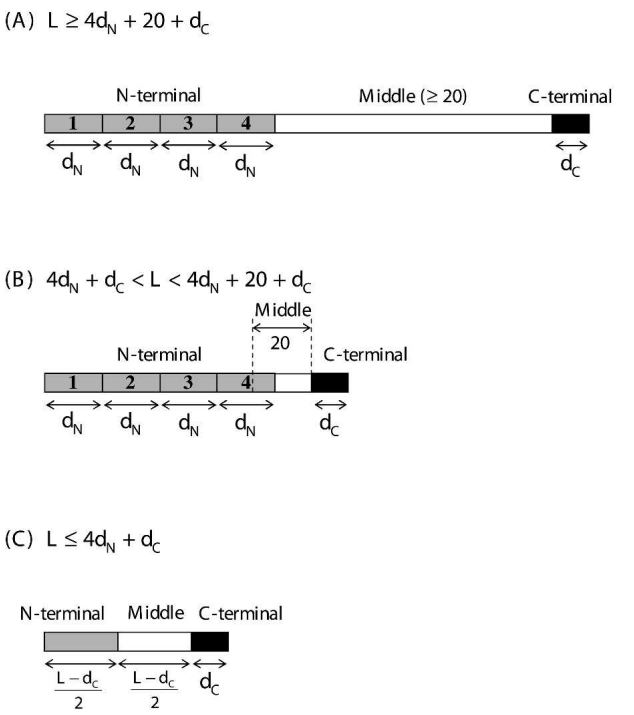

Figure 4. Definitions of the N-terminal, middle, and C-terminal parts depending on sequence length L. d_N represents the length of a region in the N-terminal part (in gray). d_C is the length of the C-terminal part (in black).

Table 8. *Regularization parameter* C *and parameter* γ *of RBF kernel used in the SVM training*

		Our method		Kim et al. (2004)	
Data set		*C*	γ	*C*	γ
NNPSL	Eukaryotic	8.0	0.64	—	—
	Prokaryotic	8.0	0.66	—	—
TargetP	Plant	5.8	0.72	10.0	0.008
	Non-plant	5.8	0.66	7.0	0.005

to location *l* (true negative). *fp*(*l*) is the number of overpredicted sequences in location *l* (false positive). *fn*(*l*) is the number of underpredicted sequences in location *l* (false negative).

Acknowledgments

We thank Dr. Bill Pearson for the help about the usage of the FASTA package and Dr. Morihiro Hayashida of Kyoto University for valuable comments. This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" and the Education and Research Organization for Genome Information Science, both from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340:** 783–795.

Bhasin, M. and Raghava, G.P.S. 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* **32:** W414–W419.

Bhasin, M., Garg, A., and Raghava, G.P.S. 2005. PSLpred: Prediction of subcellular localization of bacterial proteins. *Bioinformatics* **21:** 2522–2524.

Bruce, B.D. 2000. Chloroplast transit peptides: Structure, function and evolution. *Trends Biochem. Sci.* **10:** 440–447.

Cai, Y.-D. and Chou, K.-C. 2004. Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics* **20:** 1151–1156.

Cedano, J., Aloy, P., Pérez-Pons, J.A., and Querol, E. 1997. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266:** 594–600.

Chou, K.-C. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43:** 246–255.

Chou, K.-C. and Cai, Y.-D. 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **277:** 45765–45769.

———. 2003. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun.* **311:** 743–747.

———. 2004. Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. *J. Cell. Biochem.* **91:** 1197–1203.

Chou, K.-C. and Elrod, D.W. 1998. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem. Biophys. Res. Commun.* **252:** 63–68.

———. 1999. Protein subcellular location prediction. *Protein Eng.* **12:** 107–118.

Chou, K.-C. and Zhang, C.T. 1995. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30:** 275–349.

Cosson, P. and Letourneur, F. 1997. Coatomer (COPI)-coated vesicles: Role in intracellular transport and protein sorting. *Curr. Opin. Cell Biol.* **9:** 484–487.

Emanuelsson, O., Nielsen, H., and von Heijne, G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8:** 978–984.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300:** 1005–1016.

Guo, J., Lin, Y., and Sun, Z. 2005. A novel method for protein subcellular localization: Combining residue-couple model and SVM. *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference,* pp. 117–129. Imperial College Press, Singapore.

Horton, P. and Nakai, K. 1997. Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology,* pp. 147–152. AAAI Press, Menlo Park, CA.

Hua, S. and Sun, Z. 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17:** 721–728.

Joachims, T. 1999. Making large-scale SVM learning practical. In *Advances in kernel methods—Support vector learning* (eds. B. Schölkopf et al.), pp. 41–56. MIT Press, Cambridge, MA.

Keegstra, K. and Cline, K. 1999. Protein import and routing systems of chloroplasts. *Plant Cell* **11:** 557–570.

Kim, J.K., Raghava, G.P.S., Kim, K.S., Bang, S.Y., and Choi, S. 2004. Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine. *Proceedings of the 3rd Annual Conference of the Korean Society for Bioinformatics*, pp. 158–166. Seoul, Korea.

la Cour, T., Gupta, R., Rapacki, K., Skriver, K., Poulsen, F.M., and Brunak, S. 2003. NESbase version 1.0: A database of nuclear export signals. *Nucleic Acids Res.* **31:** 393–396.

Matthews, B.W. 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochem. Biophys. Acta* **405:** 442–451.

Miyakawa, K. and Imamura, T. 2003. Secretion of FGF-16 requires an uncleaved bipartite signal sequence. *J. Biol. Chem.* **278:** 35718–35724.

Myers, E.W. and Miller, W. 1988. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4:** 11–17.

Nakai, K. and Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14:** 897–911.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Neupert, W. 1997. Protein import into mitochondria. *Annu. Rev. Biochem.* **66:** 863–917.

Nguyen, M.N. and Rajapakse, J.C. 2003. Multi-class support vector machines for protein secondary structure prediction. *Genome Inform. Ser. Workshop Genome Inform.* **14:** 218–227.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijine, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.

Omura, T. 1998. Mitochondria-targeting sequence, a multi-role sorting sequence recognized at all steps of protein import into mitochondria. *J. Biochem.* **123:** 1010–1016.

Park, K.-J. and Kanehisa, M. 2003. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **19:** 1656–1663.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Reczko, M. and Hatzigeorgiou, A. 2004. Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics* **4:** 1591–1596.

Reinhardt, A. and Hubbard, T. 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **26:** 2230–2236.

Robinson, C., Thompson, S.J., and Woolhead, C. 2001. Multiple pathways used for the targeting of thylakoid proteins in chloroplasts. *Traffic* **2:** 245–251.

Schölkopf, B. and Smola, A.J. 2002. *Learning with kernels—Support vector machines, regularization, optimization, and beyond.* MIT Press, Cambridge, MA.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

von Heijne, G. 1990. The signal peptide. *J. Membr. Biol.* **115:** 195–201.

Yuan, Z. 1999. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* **451:** 23–26.